2013 Spring Forecast Experiment: Forecast Verification Metrics

1.) Traditional, Dichotomous (2-category) Evaluation *(excerpted from the WWRP/WGNE Joint Group on Forecast Verification Research website on Forecast Verificaiton: Issues, Methods and FAQ (http://www.cawcr.gov.au/projects/verification/)*

For dichotomous variables (e.g., precipitation/reflectivity above or below a threshold) on a grid, typically the forecasts are evaluated using a diagram like the one shown in Fig. 1. In this diagram, the area "**H**" represents the intersection between the forecast and observed areas, or the area of **Hits**; "**M**" represents the observed area that was missed by the forecast area, or the "**Misses**"; and "**F**" represents the part of the forecast that did not overlap an area of observed precipitation, or the "**False Alarm**" area. A fourth area is the area outside both the forecast and observed regions, which is often called the area of "**Correct Nulls**" or "**Correct Rejections**".

This situation can also be represented in a "contingency table" like the one shown in Table 1. In this table the entries in each "cell" represent the counts of hit, misses, false alarms, and correct rejections. The counts in this table can be used to compute a variety of traditional verification measures, described in the following sub-sections.

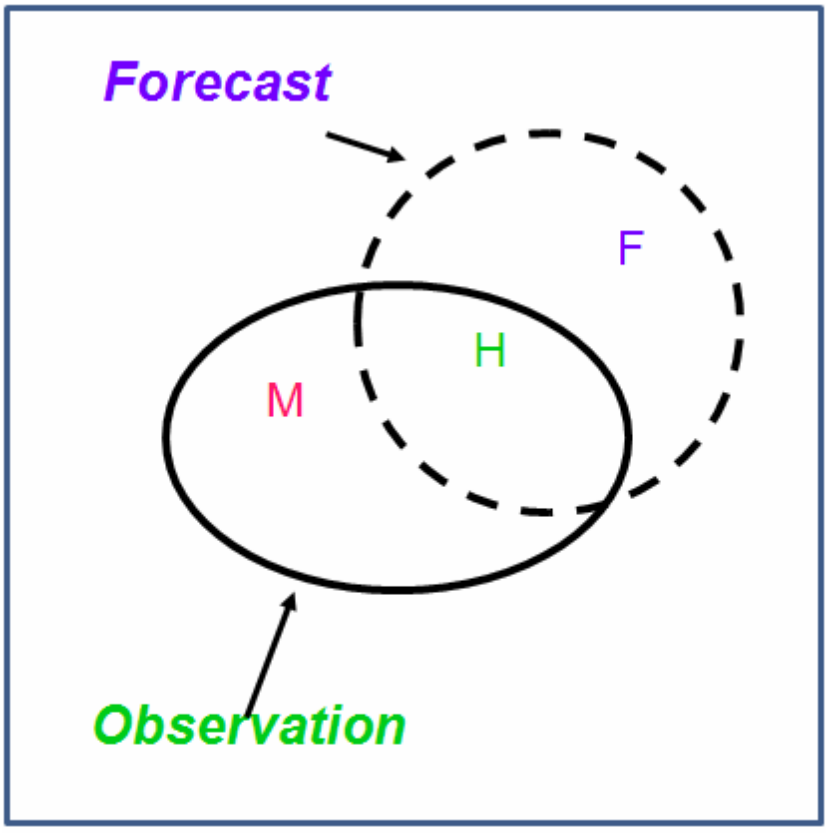

Figure 1. Diagram showing hits, misses, and false alarms for dichotomous forecast/observations.

Table 1. Contingency table illustrating the counts used in verification statistics for dichotomous (e.g., Yes/No) forecasts and observations. The values in parentheses illustrate the combination of forecast value (first digit) and observed value. For example, YN signifies a Yes forecast and and a No observation.

Forecast	**Observed**		
	Yes	**No**	
Yes	Hits (YY)	False alarms (YN)	YY + YN
No	Misses (NY)	Correct rejections (NN)	NY + NN
	YY + NY	YN + NN	Total = YY + YN + NY + NN

Probability Of Detection (POD) - $$POD = \frac{hits}{hits + misses}$$

Answers the question: *What fraction of the observed "yes" events were correctly forecast?*

Range: 0 to 1. **Perfect score:** 1.

Characteristics: Sensitive to hits, but ignores false alarms. Very sensitive to the climatological frequency of the event. Good for rare events.Can be artificially improved by issuing more "yes" forecasts to increase the number of hits. Should be used in conjunction with the false alarm ratio (below). *POD* is also an important component of the Relative Operating Characteristic (ROC) used widely for probabilistic forecasts.

False Alarm Ratio (FAR) - $$FAR = \frac{false\ alarms}{hits + false\ alarms}$$

Answers the question: *What fraction of the predicted "yes" events actually did not occur (i.e., were false alarms)?*

Range: 0 to 1. **Perfect score:** 0.

Characteristics: Sensitive to false alarms, but ignores misses. Very sensitive to the climatological frequency of the event. Should be used in conjunction with the probability of detection (above).

Critical Success Index (CSI)

Also known as **Threat Score (TS).**

CSI=TS=Hits/(Hits+Misses+False alarms)

Answers the question: *How well did the forecast "yes" events correspond to the observed "yes" events?*

Range: 0 to 1, 0 indicates no skill. **Perfect score:** 1.

Characteristics: Measures the fraction of observed and/or forecast events that were correctly predicted. It can be thought of as the *accuracy* when correct negatives have been removed from consideration. That is, CSI is only concerned with forecasts that are important (i.e., assuming that the correct rejections are not important). Sensitive to hits, penalizes both misses and false alarms. Does not distinguish the source of forecast error. Depends on climatological frequency of events (poorer scores for rarer events) since some hits can occur purely due to random chance. Non-linear function of POD and FAR. Should be used in combination with other contingency table statistics (e.g., Bias, POD, FAR).

Bias

Bias=(Hits+False alarms)/(Hits+Misses)

Answers the question: *How similar were the frequencies of Yes forecasts and Yes observations?*

Range: 0 to infinity. **Perfect score:** 1.

Characteristics: Measures the ratio of the frequency of forecast events to the frequency of observed events. Indicates whether the forecast system has a tendency to underforecast (Bias < 1) or overforecast (Bias > 1) events. Does not measure how well the forecast grid points correspond to the observed gridpoints, only measures overall relative frequencies. Can be difficult to interpret when number of Yes forecasts is much larger than number of Yes observations.

Heidke Skill Score (HSS) - $$HSS = \frac{(hits + correct\ negatives) - (expected\ correct)_{random}}{N - (expected\ correct)_{random}}$$

where

$$(expected\ correct)_{random} = \frac{1}{N}\left[\begin{array}{l}(hits + misses)(hits + false\ alarms) \quad + \\ \quad (correct\ negatives + misses)(correct\ negatives + false\ alarms)\end{array}\right]$$

Answers the question: *What was the accuracy of the forecast relative to that of random chance?*

Range: -∞ to 1, 0 indicates no skill. **Perfect score:** 1.

Characteristics: Measures the fraction of correct forecasts after eliminating those forecasts which would be correct due purely to random chance. This is a form of the generalized skill score, where the *score* in the numerator is the number of correct forecasts, and the reference forecast in this case is random chance. In meteorology, at least, random chance is usually not the best forecast to compare to - it may be better to use climatology (long-term average value) or persistence (forecast = most recent observation, i.e., no change) or some other standard.

Relative Skill (RS)

$$RSkill = \frac{CSI_{Forecast} - CSI_{MinPP}}{CSI_{MaxPP} - CSI_{MinPP}};\ RSkill < 0 \text{ when } CSI_{Forecast} < CSI_{MinPP};\ RSkill > 1 \text{ when } CSI_{Forecast} > CSI_{MaxPP}$$

$$CSI_{MaxPP} \equiv CSI_{\%} > CSI_{\%\pm1};\ CSI_{MinPP} \equiv CSI_{1\%} - \left(CSI_{2\%} - CSI_{1\%}\right)$$

Answers the question: *What is the skill of the forecast relative to a baseline reference?*

Range: Can be negative to greater than 1, 0 indicates no added skill. **Preferred Score:** > 1.

Characteristics: Used by **Hitchens et al. (2013)** to measure the performance of rare-event forecasts (i.e. SPC convective outlook slight risk areas) using "practically perfect" hindcasts **(Brooks et al. 1998)** as the baseline reference. Similar to his work, the investigation here uses CSI as the metric to compute relative skill. However, several thresholds are used to define the event object since the experimental forecasts evaluated in the Spring Experiment are probabilistic in nature and not categorical. Verification is based off of one or more local storm reports in a 40-km x 40-km grid box with a search radius of 40-km.

2.) Continuous, Probabilistic Evaluation

Fractions Skill Score (FSS)

Taken from Schwartz et al. (2010), after work by Roberts and Lean (2008)

Probabilistic forecasts are commonly evaluated with the Brier score or Brier skill score by comparing probabilistic forecasts to a dichotomous observational field. However, one can apply the neighborhood approach to the observations in the same way it is applied to model forecasts, changing the dichotomous observational field into an analogous field of observation-based fractions (or probabilities). The two sets of fraction fields (forecasts and observations) then can be compared directly. Fig. 2 shows the creation of a fraction grid for a hypothetical forecast *and* the corresponding observations. Notice that although the model does not forecast precipitation ≥q at the central grid box when the surrounding neighborhood is considered, the same probability as the observations is achieved (8/21 = 0.38). Therefore, within the context of a radius r, this model forecast is considered to be correct.

After the raw model forecast and observational fields have both been transformed into fraction grids, the fraction values of the observations and models can be directly compared. A variation on the Brier score is the Fractions Brier Score (FBS) given by

$$\mathrm{FBS} = \frac{1}{N_v}\sum_{i=1}^{N_v}\left[\mathrm{NP}_{F(i)} - \mathrm{NP}_{O(i)}\right]^2$$

where $\mathrm{NP}_{F(i)}$ and $\mathrm{NP}_{O(i)}$ are the neighborhood probabilities at the i_{th} grid box in the model forecast and observed fraction fields, respectively. Here, as objective verification only took place over the verification domain, i ranges from 1 to N_v, the number of points within the verification domain on the verification grid. Note that the FBS compares fractions with fractions and differs from the traditional Brier score only in that the observational values are allowed to *vary* between 0 and 1.

Like the Brier score, the FBS is negatively oriented—a score of 0 indicates perfect performance. A larger FBS indicates poor correspondence between the model forecasts and the observations. The worst possible (largest) FBS is achieved when there is no overlap of nonzero fractions and is given by

$$\mathrm{FBS}_{\mathrm{worst}} = \frac{1}{N_v}\left[\sum_{i=1}^{N_v}\mathrm{NP}_{F(i)}^2 + \sum_{i=1}^{N_v}\mathrm{NP}_{O(i)}^2\right]$$

On its own, the FBS does not yield much information since it is strongly dependent on the frequency of the event (i.e., grid points with zero precipitation in either the observations or model forecast can dominate the score). However, a skill score can be constructed that compares the FBS to a low-accuracy reference forecast ($\mathrm{FBS}_{\mathrm{worst}}$) and is defined as the fractions skill score (FSS):

$$\mathrm{FSS} = 1 - \frac{\mathrm{FBS}}{\mathrm{FBS}_{\mathrm{worst}}}$$

The FSS ranges from 0 to 1. A score of 1 is attained for a perfect forecast and a score of 0 indicates no skill. As *r* expands and the number of grid boxes in the neighborhood increases, the FSS improves as the observed and model probability fields are smoothed and overlap increases.

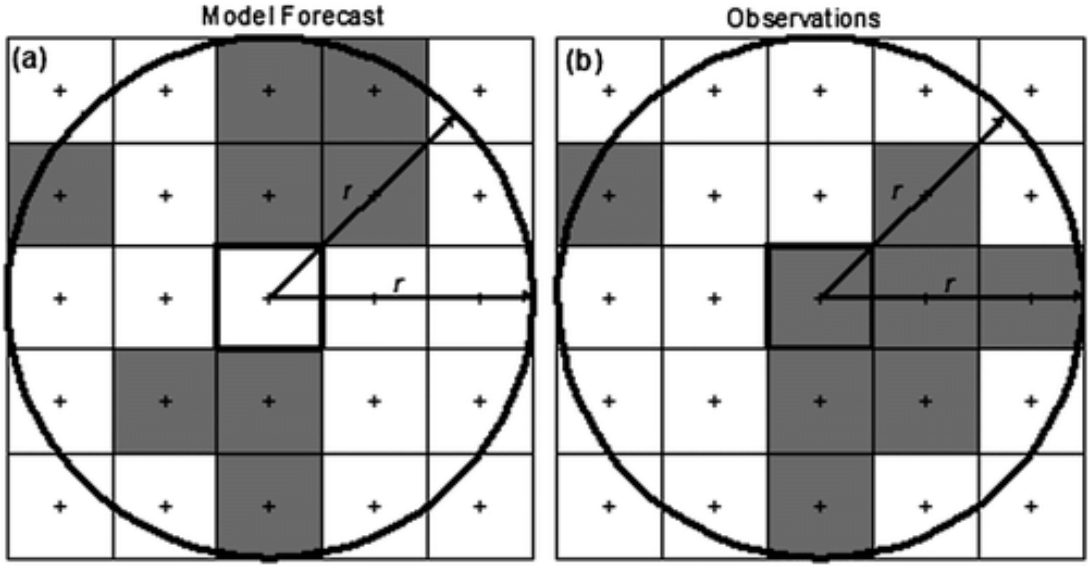

Fig. 2. Schematic example of neighborhood determination and fractional creation for (a) a model forecast and (b) the corresponding observations. The precipitation exceeds the accumulation threshold in the shaded boxes, and a radius of 2.5 times the grid length is specified.

Brooks, H. E., M. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill of rare events. Preprints, *19th Conf. on Severe Local Storms,* Minneapolis, MN, Amer. Meteor. Soc., 552–555.

Hitchens, N.M., H.E. Brooks, and M.P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev*., **136**, 78–97.

Schwartz, C., S., and co-authors: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263-280.